ARI TECHNICAL REPORT
TR-79-A11

Principles of Work Sample Testing: IV. Generalizability

by

Robert M. Guion

and

Gail H. Ironson

BOWLING GREEN STATE UNIVERSITY
Bowling Green, Ohio 43403

April 1979

Contract DAHC 19-77-C-0007

Prepared for

U.S. ARMY RESEARCH INSTITUTE
for the BEHAVIORAL and SOCIAL SCIENCES
5001 Eisenhower Avenue
Alexandria, Virginia 22333

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director

WILLIAM L. HAUSER
Colonel, US Army
Commander

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: TR-79-A11	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): PRINCIPLES OF WORK SAMPLE TESTING. IV. GENERALIZABILITY,	5. TYPE OF REPORT & PERIOD COVERED: Final rept., 15 Nov 76 - 15 Jun 78
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Robert M. Guion; Gail H. Ironson	8. CONTRACT OR GRANT NUMBER(s): DAHC19-77-C-0007
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Bowling Green State University, Bowling Green, Ohio 43403	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q161102B74F
11. CONTROLLING OFFICE NAME AND ADDRESS: US Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333	12. REPORT DATE: April 1979 13. NUMBER OF PAGES: 29
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): --	15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

Monitored by G. Gary Boycan, Engagement Simulation Technical Area, Army Research Institute

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Measurement theory, psychometrics, work sample testing, validity, content-referenced testing, criterion-referenced testing, generalizability theory

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Three kinds of generalizability are described: generalizability of test content, generalizability of relationships such as are described by validity coefficients, and generalizability of test scores across varying conditions. All are deemed important in the construction and evaluation of work sample tests. A special problem is determining the limits of generalizability; it is suggested that the research designs of generalizability theory can be applied in some cases to determining the limits of the generalizability of criterion-related validities.

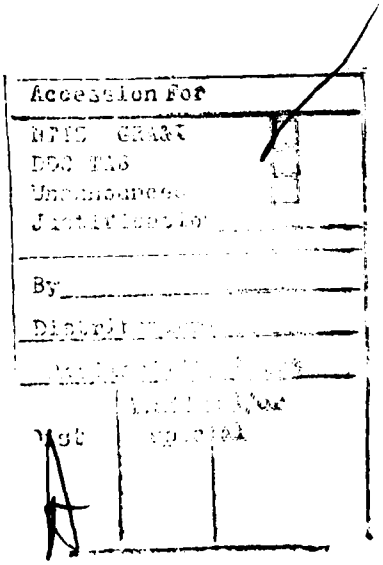
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (continued)

This report, the last of four, is written for psychologists and others interested in theories of testing.



PRINCIPLES OF WORK SAMPLE TESTING: IV. GENERALIZABILITY

BRIEF

Generalizability has been identified as a major concern in the evaluation of work sample tests. Generalizability is also an ambiguous term that has had substantially different meanings in different contexts. It sometimes refers to generalizability of test content to a broader test domain. It has often been used in the context of validity generalization, a special case of the generalizability of relationships. It has been recently used frequently to refer to generalizability theory, or an approach to the generalizability or dependability of scores.

An example of systematic effort to assure generalizability of content is cited from the development of a model tank gunnery test (Wheaton, Fingerman, & Boycan, 1977); this work is cited particularly for the development of an index of generalizability. Its applicability is particularly great for direct work samples. For abstracted work samples, special attention should also be given to the importance of specific task variables and to matching these variables in the construction of the abstracted sample.

Generalizability of relationships, as in the testing of a criterion-related hypothesis, is understood best in the example of the generalizability of criterion-related validities, but there are other examples. It is a criterion-related hypothesis that needs to be tested and, if supported, evaluated for its generalizability when a literal measurement of one variable is used for inferring measuring of a different variable; this is usually done under the rubric of construct validity. A similar example is the abstracted work sample, in which the literal performance of the unreal task is used for inferring performance on the real one. Studies of generalizability of relationships need to pay particular attention to the limits of such generalizability.

Generalizability theory, as proposed by Cronbach, Gleser, Nanda, and Rajaratnam (1972) refers to the generalizability of scores. Several designs for such studies are described. A general conclusion is reached that the limits of validity generalization (i.e., of the generalizability of relationships) might be studied by the generalizability theory research designs.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

Each of the preceding reports of this series has concluded that a major consideration in the evaluation of a work sample test is its generalizability. Unfortunately, the word is so ambiguous that the conclusion can mean different things to different readers. Consider the following list of statements drawn primarily from the other three reports:

1. In an experiment for evaluating programs or materiel, the sample of performance used as the dependent variable of the experiment should be representative of -- that is, should generalize to -- the performance in a target situation.

2. The properties of the tasks included in a work sample should be representative of the properties of the tasks in the domain from which they are sampled; for example, if, in the job content domain the job is structured so that each batch of 20 completed assemblies is a recorded unit, then the work sample should be based on a similar work unit. To the extent it does not, performance on the work sample may not generalize to performance on the real job.

3. The first step in the evaluation of a work sample is to evaluate the degree to which it is congruent with the domain being sampled; are the salient task characteristics adequately sampled in appropriate proportions?

4. The simple, abstracted work sample is best evaluated in terms of how well performance on the abstraction can be used to infer performance on a more highly complex real job. Performance on the abstract portion must generalize to performance on the whole.

5. In measurement in institutional settings, the actual measurements obtained usually are of less concern than are attributes to be inferred or predicted from them; the evaluation of measurement under conditions of institutional control is based on the degree to which the scores will generalize to attributes of greater institutional concern.

6. Measurement of one attribute is often inferred from literal measurement of quite a different attribute, such as the use of reaction time to infer some aspect of information processing. The usefulness of such an arrangement depends on how well performance on the one attribute generalizes to performance on the one inferred.

7. The criterion-related validity obtained in one setting may generalize to a different setting if the characteristics of the subject populations, tasks, and settings are essentially similar.

8. A disadvantage of tight, experimental control of measurement in a laboratory situation is that the situation may differ so substantially from the target situation that measurement in the one will not generalize to the other; where the situations differ markedly, "generalizability must be assessed."

9. The evaluation of work samples in field settings is incomplete unless it can be shown how well performance in the field setting used will generalize to performance in some targeted setting.

10. The variables that describe a task -- autonomy, demand for unwavering attention or other skills, amount of feedback, and others -- may change systematically as characteristics of the situations in which performance is measured change; as task variables change, generalizability may be reduced.

All of these statements are special instances of the broader statement that, regardless of purpose, setting, or combination of variables and methods of measurement, generalizability is the universally required characteristic of effective measurement. They do not, however, merely say the same thing in ten different ways. The first three statements refer to the generalizability of the content of the measurement; they deal broadly with the generalizability of the actual content of the content sample chosen to the rest of the domain that might have been chosen. The next four statements all deal with the generalizability of a relationship; they ask about the relationship of the variable measured to a variable one may perhaps wish had been measured, and, in statement 7, about the generalizability of such a relationship

across settings. The last three statements all refer explicitly to the generalizability of scores, or of inferences from scores, across different conditions of measurement.

In almost every one of the statements in that list, there is a sense in which all three kinds of generalizability may be relevant; nevertheless, these three can be examined independently both for the concepts they represent and for the possible methods available for evaluating them.

GENERALIZABILITY OF CONTENT

Every sampling is intended to generalize to the whole from which it is a sample. Random sampling in experimental work is intended to assure generalizability of results to the population sampled; stratified sampling is the effort of people whose faith in random sampling is small to assure such generalizability.

So it is with content sampling in the development of a work sample. Each stage in the process is a sampling exercise, not necessarily representative, which defines the part or aspect of the previous whole to which conclusions of some sort are expected to generalize. Job content domains are intended to generalize to certain aspects of job content universes; so it is also with test content domains and universes, and a test content domain is expected to have potential for generalizing to selected aspects of a job content domain. And the sample of the test content domain ultimately chosen is certainly expected to generalize to the test content domain; that is what so-called content validity is about. The problem is the choice of procedures by which one may feel comfortable in making generalizations from the sample to the whole.

A CASE STUDY IN ITEM SAMPLING

Wheaton, Fingerman, and Boycan (1978) dealt explicitly with the generalizability problem in item sampling and developed an index of generalizability to be applied to "items" of a work sample. The work was that of a tank crew in "neutralizing" a target; the items were job objectives (of which they identified 266). Each job objective consisted of some subset of a possible 114 behavioral elements. By cluster analysis of a matrix of the 266 objectives and 114 elements, fourteen empirical and two rational clusters of objectives, or families, were identified. (In the third report of this series, the steps recommended for job analysis follow a similar pattern; to compare the two reports, job objectives can be equated with the component tasks of report III; behavioral elements to task elements; and families to task categories.)

The procedure followed was related by Wheaton et al. to the classical procedures of item analysis; they argued (a) that classical item analysis produced a set of items for a final test that would generalize to the whole and (b) that classical item analysis was too expensive to duplicate in a tank gunnery test. The present report will not follow the first of these arguments; what they did seems preferable to the more expensive item analysis. Empirical item analysis techniques will succeed in selecting a subset of all original items that will maximize the overall variance in that set. That is, a skillfully conducted item analysis, with certain sets of data, will result in the elimination of those items which have a low correlation with the composite of all other items, and the resulting test will have a variance comparable to that of any other similar-size subset of items and perhaps even a reliability as good as the total set.

Items eliminated, however, may be eliminated only because they are low in variance in the particular sample in which the item analysis is done. Alternatively, the items may be eliminated because they have low common variance, although substantial variance of their own, and therefore do not contribute to the internal consistency of the final test.

The argument of these reports is that high levels of internal consistency are not necessary in work sample tests, even though subscores on a work sample should have enough internal consistency to demonstrate some degree of functional unity. If this argument is accepted, it follows that an item need not be eliminated from the test (or subtest) unless it actually shows a zero or negative correlation with the rest of the test. Thus an empirical item analysis might have eliminated some items identified as important by the procedures used by Wheaton et al. -- items which helped assure that the final test was indeed representative of the behaviors actually required on the job.

The rational foundation for the approach is the assumption that the more behavioral elements that one job objective has in common with other job objectives, the more generalizable is the performance on it relative to the others. Arbitrarily but correctly, they believed that each family of job objectives should be represented in the final work sample according to the number of objectives it contained; that is, a family or cluster of objectives with 30 objectives in it should be represented three times as heavily as a ten-objective cluster. The issue was to choose the one or the three objectives as items of the work sample so as to maximize the generality of the resulting test.

Several methods were considered and rejected. For example, one

might take the objectives with the largest number of behavioral elements. This was rejected because many of the elements might occur only in that one objective, that is, they might not generalize to the family of objectives.

Two methods were considered intuitively sound and were ultimately combined. Each of them is a method of obtaining a weighted sum of the behavioral elements included in each job objective. One method of weighting assigns a weight equal to the frequency that element i occurs within a family, that is, F_i. This assigns the greatest weight to the most commonly occurring elements. The second method is to assign a weight of F_i/D_i, where D_i is the frequency with which the element occurs in the total domain; this method would give the greatest weight to an element which occurs almost exclusively within the one family, even though it may not appear very often. The two methods were combined by multiplying them so that the weight recommended by Wheaton et al. is $\Sigma(F_i^2/D_i)$, called the index of generalizability, for each job objective. Although their decisions were based in part on practical issues of cost as well as on these indices, a relatively inexpensive content domain could be sampled largely empirically through the use of such an index.

ABSTRACTED TESTS

The item sampling reported by Wheaton et al. is an example of the direct work sample; that is, the job objectives chosen were component parts of the actual jobs selected because they were more widely representative than others. In the abstracted work sample, the "items" -- or at least the composites which the set of items produce -- may be created rather than selected by sampling. The abstracted work sample identifies certain components -- task elements -- and puts them

together in such a way as to represent, if not to resemble, the essential or crucial elements of the job itself. Again, the purpose is to select a set of final tasks that will generalize to the whole of the original content domain, even if less intuitively obviously.

The task elements might be chosen by a procedure not unlike that for weighting elements in the Wheaton et al. example. An alternative is to have a panel of qualified judges (i.e., qualified because of their knowledge of the job) rate each element both for frequency and importance. Although these ratings will probably be highly correlated, they are both necessary if the task elements that occur rarely but are crucial when they occur are to be identified (or conversely, if there is to be any success in identifying the task element that is important because it is pervasive.) Some combination of these ratings can be used as the basis for selecting the task elements to be abstracted that will generalize most handily to the domain, at least as the term is used by Wheaton et al. There remains, however, the problem of putting these elements together into a task so that performance on the artificial, abstracted task will generalize to performance on the real job.

At the outset, it should be clear that this poses an empirical, not simply a rational problem. If little abstraction is involved, there may be no reasonable basis to question the relevance of the abstraction to the job as a whole; if the abstraction is severe, however, the relationship between performance on it and performance on the job itself is subject to investigation by empirical study.

Some things can be done in the test development stage, however, that will either minimize the need to do the empirical study or maximize its probability of success. These are things that maximize

the basic similarities between the real tasks and the artificial abstracted tasks.

The first report of this series, the taxonomy of testing, reported nine possible categories of task variables. These included;

1. Duration or intensity of attention
2. Hazards
3. Degree of task structure
4. Organizational involvement
5. Task complexity
6. Intrinsic feedback
7. Skill demands
8. Significance
9. Autonomy

It would appear intuitively that attempts to match the two tasks on the most salient of these kinds of variables would assure generalizability of the abstraction to the domain.

INDUSTRY-WIDE RESEARCH

There seems to be growing concern for developing research programs to validate employment tests on an industry-wide basis: the banking industry, the insurance industry, and the electric power industry have all initiated such research. In such studies, the purpose is to develop tests that will generalize across various organizations which, independently, make up the industry. In part, this

discussion belongs under the next major heading, the generalizability of relationships, since the major target is to develop criterion-related validity statements that will generalize from one company to another. However, before that portion of the objective can be reached, it is first necessary to define the job domains to be tested. Different jobs in different organizations may involve the same total collection of component tasks put together in different ways. The problem is to identify a subset of component tasks which will themselves have generality, which is to say, will generalize, across organizations.

GENERALIZABILITY OF RELATIONSHIPS

Every criterion-related hypothesis is an example of a generalization; the hypothesis, if supported, means that inferences about a criterion can be drawn (or generalized) from performance on a predictor. If the hypothesis is supported often enough, and most of the time, the validity of the hypothesis is itself generalized to situations beyond that in which it was initially supported.

THE CRITERION-RELATED HYPOTHESIS

The most commonly proposed hypothesis in personnel testing is the predictive hypothesis that performance on an employment test can be used in a functional equation to predict performance on the job. Variations on the theme include the hypothesis that a work sample can be used, perhaps in something like a discriminant function, to classify examinees correctly into master and non-master categories defined on some basis other than the test. In such examples, it is clear that the predictor, the test score variable, and the criterion, a job performance variable, are measuring quite different classes of

attributes. The fact of different variables is what sets the investigation of the criterion-related hypothesis apart from investigations of test validity which inquire into the excellence of the test as a measure of some variable.

The list of statements at the outset in this report refers to the conventional criterion-related hypothesis in statement 5; perhaps this statement describes the essential criterion-related hypothesis for personnel testing: the hypothesis that the ordering of people according to their test scores can generalize to their order according to a variable "of greater institutional concern." The essence of the criterion is that it is the variable of greater organizational interest. Whenever scores on a test are used not as descriptions of test performance but as indicators of something else that is more important, a criterion-related hypothesis is at least implied. Two kinds of implied hypotheses need to be identified, the hypothesis implied by using one variable as a presumed measure of a different variable, and the specific example of that presumption in work sample testing where an abstraction of the job is a presumed measure of performance on the real thing.

Inferences from Literal Measures. Psychological measurement typically measures one thing for the purposes of inferring something of greater interest. The number of arithmetic problems correctly solved in a specified time period is taken as a sign of an underlying construct called numerical fluency. Correct identification of where small areas on a flat surface would be found if the surface were folded into three dimensional objects is used at one level as a measure of an underlying ability called visualization and, at another, of a construct called mechanical aptitude. The example used earlier is of measuring the electrical resistance on the surface of the skin

and, as it changes, inferring changes in emotionality.

These are criterion-related hypotheses, but they are not among the examples of criterion-related validation in conventional approaches to personnel testing. Rather, the evaluation of these hypotheses proceeds along the lines known as construct validation; if a substantial body of research, consisting largely of concurrent criterion-related validity studies testing these hypotheses, supports these presumptions, then the literal measure is said to be valid for measuring (or, more properly, inferring) the more interesting measure. In a dictionary sense of the term, such an inference is a generalization from the narrowly defined literal measure to the broader meaning of the variable of greater interest. Such generalization nearly always requires some form of empirical support, although it is not always known as construct validity. In experimental psychology, particular measures are generalized because of the weight of the literature using them. If a set of operations for defining a construct becomes widely enough accepted, the issue of construct validity in its psychometric sense is not raised. Thus, if there is enough acceptance of performance on a specific set of tasks, none of which is genuinely drawn from a job, as an indication of performance on a job these tasks were designed to reflect, the construct validity of the interpretation of the performance as a general ability is not questioned.

Abstracted Work Samples. That is to say, the abstraction that occurs in the development of a work sample test may be accepted without raising questions of construct validity. The requirements are that informed people understand the nature of the abstraction and its relevance to the actual job or that a history of its use has demonstrated empirically that relevance. The latter is unlikely; unlike

experiments in basic psychology, the development of an abstracted work sample is, unless the level of abstraction is extreme, an ad hoc affair.

The nature of the abstraction and of its relevance may be self-evident to informed people if the degree of abstraction is not great. For example, it is unlikely that any informed and reasonable critic would object to a work sample test for cabinet makers that resulted in no useful product but did manage to incorporate nearly all of the most difficult kinds of joining, routing, and other skills. Real questions could be expected, however, if a test of steadiness in running a piece through the same motions used in rabbeting were used in place of cutting an actual rabbet. Where these questions arise, the developer of the test has, knowingly or not, proposed a criterion-related hypothesis in which the criterion is actual job performance. The purpose of the test is to generalize to inferences about the construct of job skill. Its construct validity may be supported in part by the logic of the content sampling leading to the abstraction, but it is better supported by empirical evidence that the generalization can legitimately be made.

VALIDITY GENERALIZATION

Assuming that the criterion-related generalization has been empirically supported, the next question that arises is whether that support holds only in specific settings. This is the question of validity generalization. Lawshe (1952) spoke of generalizable criterion-related validities as being those, like typing tests, which had found so much support that situational validity studies were not needed.

Conventional wisdom has long asserted that criterion-related validities need to be determined independently in each situation of application. The wide spread of obtained validity coefficients, such as those described by Ghiselli (1966) was taken as evidence of the need to determine for each situation the validity of the hypothesis in that situation since validities vary so widely. The present writer has noticed an embarrassing example of the hypocrisy of the demand for situational validation. In one paragraph (Guion, 1965, p. 455), it was pointed out that personnel testers have no scientific generalizations on which to depend. The next paragraph identified as the first of three "serious flaws" in selection research the tendency of personnel testers to stick with the "safe" tests -- those for which validity had been repeatedly demonstrated.

The Schmidt and Hunter Study. Schmidt and Hunter (1977) noted that many factors may be responsible for obtaining different validity coefficients in specific samples drawn from a common population. If, for example, the population consists of all people who work at a specific job tested by a specific type of test, then there is a true population validity coefficient. Obtained validity coefficients from finite (and usually rather small) samples will be distributed around that value. By making two assumptions about criterion reliability and about degree of restriction of range in these sample -- assumptions which seem not to be unreasonable -- Schmidt and Hunter determined the probable standard deviation of the sampling distribution. Their conclusion was that the variability of validity coefficients reported in various reviews can be almost entirely explained in terms of these two statistical characteristics of the samples used. When one adds further artifactual influences on obtained validity coefficients, such as differences in factor structure of individual tests within the test type or of criterion measures in individual studies, the room for

real situational variables accounting for differences in obtained validity coefficients is slim indeed.

Limits of Generalizability. The good news according to Schmidt and Hunter should not be prematurely embraced as evidence of validity generalization as the invariant rule. The logic of much of the field of industrial and organizational psychology argues against such an interpretation. Attempts to improve training will, for example, vary widely from one organization to another so that the effectiveness of job training will account for different portions of criterion variance. Organizational development programs, human relations training for supervisors, and attempts to provide or manipulate reward systems are all psychological attempts to account for substantial portions of performance variables, and the existence and success of such attempts will vary widely. Nonpsychological influences on performance should account for varying portions of the criterion variance; equipment in some organizations is the most advanced and sophisticated available; in others, it is said to be held together with chewing gum and baling wire. It is not reasonable to assume that the same "true" validity coefficient applies to the organization with the best leadership, training, and equipment as to the organization where no advances have been made in these fields.

Yet the argument of the Schmidt and Hunter analysis is persuasive; evidence can be found of the generalizability of observed relationships between predictors and the criteria they predict.

The resolution of the apparent contradiction seems obvious enough; the question is not whether relationships generalize beyond the situation in which they are first observed, but rather how far -- within what limits -- may that generalizability be assumed.

This question has not been addressed in the literature reviewed to date. Work in progress under a Fellowship at Bowling Green State University may point to one approach to it. It is currently concerned with the generalizability of various relationships at the management level. As preliminary research, three taxonomies are being developed: (a) a taxonomy of management occupations, (b) a taxonomy of management styles, and (c) a taxonomy of situations. As the work progresses, it will try out certain relationships rather well established in the research literature in each of several of the cells of the resulting three-dimensional matrix. If variability is (a) random and (b) explainable in terms of artifacts such as those described by Schmidt and Hunter, then generalizability across the domain may be assumed. If variability is (a) random and (b) greater than can be explained by statistical artifacts, then influences on validity exist which have not been included in this model. If variability is, however, substantial and fitting a pattern, then the three-dimensional model may be useful in identifying the reasonable limits within which the relationship may be said to generalize.

The taxonomic approach has serious flaws, not the least of which is the problem of arranging the cells for identifying patterns of obtained coefficients. It is, however, one approach to the search for limits of generalizability, and others need to be generated.

GENERALIZABILITY OF SCORES

The most completely described approach to generalizability is that of Cronbach, Gleser, Nanda, and Rajaratnam (1972) applying analysis of variance research designs to the study of the dependability of scores. Although the method and theory so far promulgated refer only to the generalizability of scores, they may have implications for

modifications that can help identify the limits of validity generalization as well.

Generalizability theory extends classical psychometric theory in several ways (Brennan, 1977; Cronbach et al., 1972). The most important of these is that generalizability theory recognizes distinct sources of error in any measurement, in contrast to a single, undifferentiated source of error. In classical test theory, reliability is defined as the proportion of observed score variance attributable to "true" score variance (that is, not random error variance). Generalizability theory replaces the reliability coefficient with the coefficient of generalizability, the "true" score with the more precisely delimited universe score, and random error variance in general with specific sources of error variance.

The definition of a universe depends on how the investigator plans to interpret the measure. Cronbach et al. (1972) identify three possible kinds of decisions; classification of people on the basis of scores into two or more categories, comparisons of possible courses of action for the same persons, or normative comparisons of people. The defined universe of generalization depends pretty much on the kinds of decisions. For some of these it may be the entire universe of admissible observations. For others, it may be a subset. A third possibility, that it may be something outside of that universe, has been rejected by Cronbach and his associates as either referring to validity or to an overgeneralization.

The universe score is the expected value of the observed score for a person over the universe of items. The coefficient of generalizability is the ratio of universe score variance to expected observed score variance. Clearly, there may be several coefficients of

generalizability; each of these generalizability coefficients explicates the domain to which the coefficients of generalizability are applicable.

Perhaps the major advantage of generalizability theory is that it makes the investigator ask questions which might otherwise not be considered (Cronbach, 1976). That is, the investigator must make an explicit choice of the nature of the universe to which scores are expected to generalize. The system then dictates, at least in large measure, the plan or design for data collection and analysis. The proper design depends on the nature of the universe chosen.

Replacing "true score" with "universe score" emphasizes that the investigator is making inferences from a sample of possible observations; the choice of universe emphasizes that there is more than one universe to which an investigator might wish to generalize. To be more explicit, if a person is tested on a work sample, he may himself be considered as belonging to a population of males, Californians, whites, or others; he may be tested in a population of circumstances, involving different samples of tasks, or time of day, or level of environmental hostility, or by different observers. Thus, in generalizability theory, any measurement of an attribute is considered to be a sample from some large set or universe of possible measurements, the universe of admissible observations, defined by observations in all possible combinations of circumstances.

Circumstances of the same kind are called facets. ("Facets" is chosen in preference to "factors" -- a more common term in analysis of variance -- to avoid confusion among testers with the factors of factor analysis research.) Within the universe of admissible observations, each facet consists of two or more categories or conditions.

Facets in the example above might include observers or task difficulty levels or test environments; still another might be race of the person. A decision maker may wish to generalize the results of measurement to only a limited portion of the overall universe defined by these facets; that portion is the universe of generalization. For example, a study of supervisory ratings might (conceivably) include as facets time of day of rating, nature of assignments given to the subordinate who is to be rated, race of either supervisor or subordinate or both, and height of supervisor. For particular problems, a decision made might wish to generalize only across assignments and race and ignore height or time of day.

In designing a study, all facets that might influence scores should be identified. Cronbach et al. (1972) distinguished between a generalizability study (a G study) and a decision study (a D study). The G study is designed to estimate components of variance attributable to different facets among all those that might account for substantial portions of variance; the G study therefore defines the universe of admissible observations broadly. A variety of D studies might follow from this broad definition, each of which uses whatever information is relevant to the decision maker. The distinction between the two types of studies is conventional but not very useful in this discussion of work sample testing. All work sample testing is done for the purpose of making a decision (even, if the concept of decision is broad enough, when the test is to be used as the criterion in the criterion-related validity study of some other predictor). The facets that influence work sample scores are those that indeed influence the decisions. If, for example, an abstracted work sample does not generalize from training center to field conditions, but a direct work sample does, then the decision in determining a candidate's qualifications for the field job will be based on the latter. The

discussion that follows assumes that the universe of generalizability for general evaluation of the measurement is not appreciably different from that in making decisions either about individuals or about tests, and the distinction will not be maintained.

SOURCES OF ERROR IN WORK SAMPLES

Glaser and Klaus (1962) discussed six sources of error influencing the reliability of performance measures:

1. The test environment varies; a driver's test for licensing, for example, may be given during good weather, rain, or snow.

2. System instability may cause unreliability; fluctuations in the condition of equipment (for example, wet or dry brakes) or in the behavior of other people in the system (for example, other drivers on the same freeway) influence performance.

3. The equipment may be different; in the above example, different people drive different automobiles of different ages and different systems for shifting gears.

4. The sampling of tasks may vary. The tasks in a driver's examination in a major metropolitan area are different from those in a rural county. In one state, candidates may choose between parallel parking and driving a zig-zag obstacle course.

5. Dimensions describing the complexity of the behavior required by the task may vary; they should be consistent.

6. Examinees reactions vary from time to time. Part of this is pure random variation, but part of it may be due to differences in conditions at different times, for example, when the examinee is closely observed or when the examiner is of a different race.

Since work sample tests are often scored by the judge's evaluation, a single reliability estimate is not adequate. At the very least, one needs to evaluate the reliability of the observer and the overall test reliability (Lang, 1978). A major advantage of generalizability theory

is that one can design a study to estimate the contribution to error that each of these different sources makes and, moreover, to find a generalizability coefficient which specifically accounts for whichever source of error is of particular interest. For example, one can find a coefficient of generalizability which describes whether the evaluation of performance on a work sample generalizes over different observers, over different conditions of testing, or different specific items in the work sample. In addition, a generalizability study would help an investigator decide how many judges are necessary to get dependable evaluations, or how many conditions of testing might be necessary to obtain a satisfactory generalizability coefficient.

DESIGNS

Consider a simple design. A sample of items is presented to a sample of people. In this design, using analysis of variance terminology, items and people are crossed. This is represented by the Venn diagram in Figure 1. The classical estimate of error, $\sigma^2(\delta)$, includes only the variance due to the center section of the diagram. That is, the portion of the total variances due to persons and to items interacting. Generalizability theory introduces a different error term, $\sigma^2(\Delta)$, which incorporates the additional error due to item sampling (see Cronbach et al., 1972, p. 24). Here the person is the object of the measurement and we may wish to generalize over items. Brennan (1977) presented formulas for the computations and discussed the design in greater depth. An important distinction which he made related to the appropriateness of each of these in content-referenced tests. He pointed out that, whereas $\sigma^2(\delta)$ is a measure of the variance of relative error, $\sigma^2(\Delta)$ is a measure of the variance of absolute error and is therefore much more appropriate for use in mastery testing and other forms of content-referenced measurement.

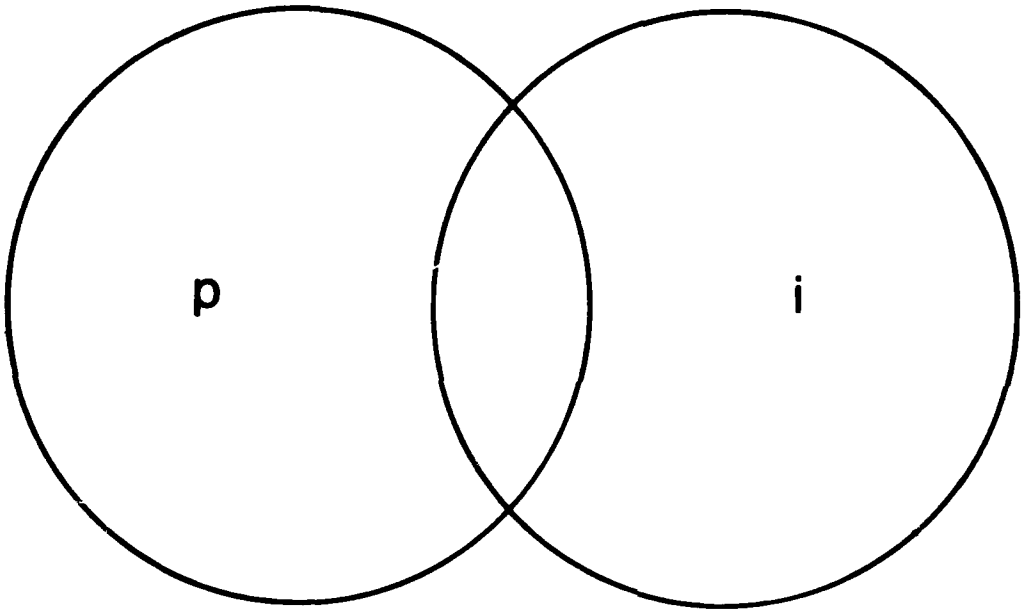


Figure 1. Venn diagram identifying variance components for estimating generalizability of scores of persons. (p = persons; i = items; area of overlap identifies p x i interaction, or error)

Dependability of Instructor Ratings. In this section, two more complicated designs which have been used in evaluating student ratings of instruction can be discussed in the context of peer ratings or observer ratings used in evaluating work samples.

The key issue in a study reported by Kane, Gillmore, and Crooks (1976) was the dependability of the mean rating of an instructor over a class of students and a set of items. The design was basically a split plot design; students were nested within classes and crossed with items, as illustrated in Figure 2. Figure 2 identifies five sources of variance: class, students within the class (confounding the student mean effect and the student-by-class interaction), items, interaction of items and class, and the residual (which includes the student-by-item interaction, the student-by-class-by-item interaction, and error). These sources of variance essentially define the nature of a generalizability study.

This may be placed in a context of peer ratings of trainee performance in a sample of work. The sample of work is done in different settings. Consider it to be a total job sample observed for a period of two days under the different conditions. The conditions might be night and day. After the two-day period, those peers who have had contact with the trainee might rate his performance on a set of rating scale items. The design for analysis is judges or peers nested within condition and crossed with items. We are interested in the dependability of peer ratings of the trainees. The choice of which of three generalizability coefficients is appropriate depends on whether one wishes to generalize over judges and items or only over judges or only over items. Explicit formulas for computing these generalizability coefficients are given by Kane et al. (1976).

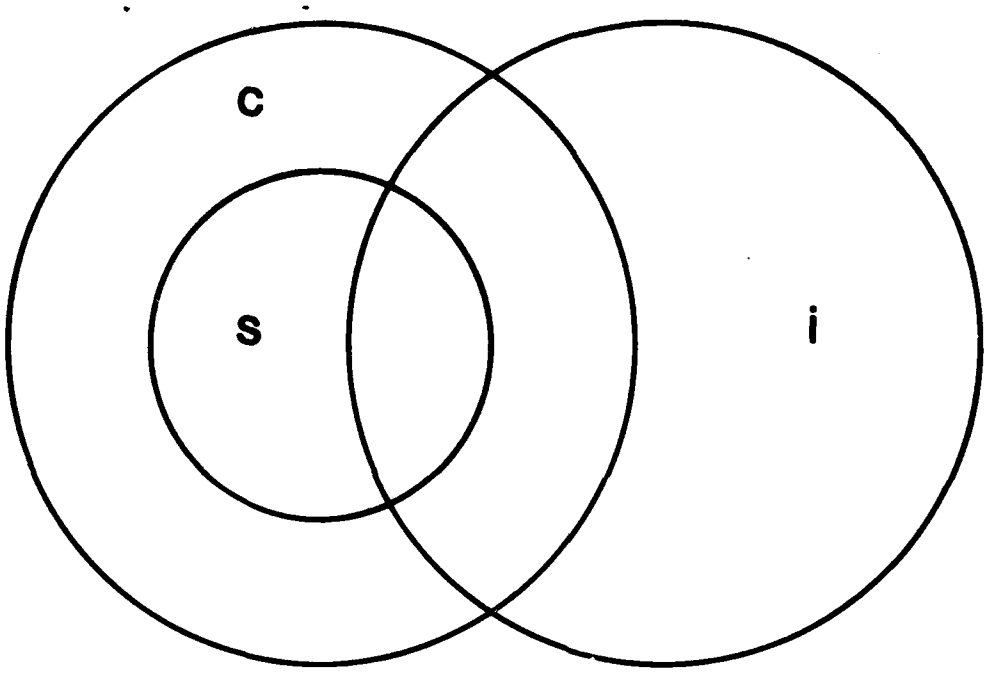


Figure 2. Venn diagram estimating variance components in student ratings of instructors. (c = class; s = students; i = items of scale; areas of overlapping circles identify various interaction terms)

A study such as this provides not only the three estimates of generalizability, but it would provide enough information so that the number of judges and the number of items could be determined for the desired size of the generalizability coefficient. In the example involving student evaluation of teaching, the authors recommended using 15 or more students and four or more items for what they considered to be a satisfactory generalizability coefficient over both students and items -- roughly .70. Moreover, it was found that increasing the number of students had a greater impact on the generalizability coefficient than increasing the number of items. In the analogy of the peer ratings of trainees, this would mean that increasing the number of peers doing the rating would have a greater effect than increasing the number of items on which the rating is done. This seems intuitively logical on the basis of earlier work showing that classical reliabilities increase as a function of the number of raters in accordance with the Spearman-Brown prophecy formula.

Several questions might be asked that this generalizability design would not answer. Differences in the ratings of trainees are confounded with conditions. If one trainee is placed in a night condition and another in a day condition, differences in the ratings might be due either to differences in the trainees or to differences in the conditions. In the school example, this is a confounding of teacher and course effects.

It is possible to design a study which would separate out the trainee effect from the effect of conditions. Perhaps it might be useful to know the variety of conditions over which performance of the trainee is expected to generalize. The obvious solution is to have a number of trainees rated in each of several conditions. Such a completely crossed design may often not be feasible in practice. If

not, one might design the two studies pictured in Figure 3. In the first of these, conditions are nested within trainees, thereby providing an estimate of the variance component for the main effect due to trainees, and there is also an estimate of the variance component for the condition main effect confounded with the condition by trainee interaction. One can estimate the generalizability of the average rating of the trainee over conditions from this study. A different study, nesting trainees under condition, will give the variance component due to the main effect of conditions and an independent estimate of the variance component for the trainee main effect confounded with the trainee by condition interaction. This study will give an estimate of the generalizability or dependability of the average rating of a condition, generalizing over trainees. Computational details for this design applied in an educational setting have been reported by Gillmore, Kane, and Naccarato (1978).

For the first study in Figure 3, there are $2^3 = 8$ possible combinations of facets in the universe of admissible operations. We can generalize or not generalize over conditions, raters, and items when the object of measurement is the trainee. One can determine how the generalizability coefficients will change when the number of conditions or raters or items is changed. The logic involved in analyzing the second of these studies is similar.

Instead of conditions (night vs. day), the condition facet might have included different field conditions ranging from highly supportive, as in the training situation, to exceedingly hostile, as in realistic maneuvers or combat. A different facet might be region or part of the country or theater of operation, another one might be the nature of terrain, still another might be organizational characteristics, such as structure or "climate." In designing a generalizability study,

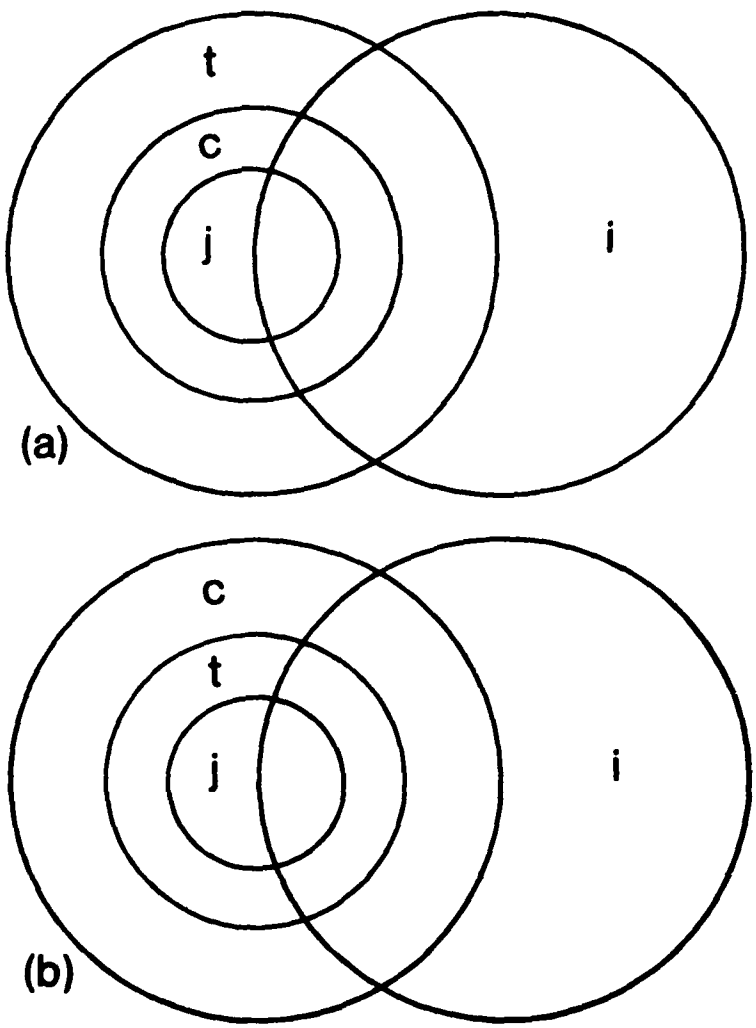


Figure 3. Two generalizability designs for estimating trainee and condition components. (t = trainees; c = conditions; j = judges or raters; i = items of rating scale)

one basically needs to identify precisely what it is that is being measured (for example, peer ratings of trainees) and the facets to which that measurement needs to be generalized. Brennan (1977) clearly covers specific details for five common designs. Cronbach et al. (1972) identifies a broader range of possible designs.

SUMMARY

The ambiguous nature of the term, generalizability, when examined leads to interesting and important research on the evaluation of work sample tests. Three meanings of the term have been identified: Generalizability of content to a broader domain, generalizability of relationships as in criterion-related validities, and generalizability of scores over conditions of measurement.

Cronbach and his associates have explicitly excluded traditional concerns for criterion-related validity from their notion of generalizability theory. A form of construct validity is demonstrated when a measure such as a work sample, which is defined primarily as a sample of job observations, is found to generalize over important varieties of conditions.

A consideration of the Schmidt and Hunter (1977) work suggests a further possibility for applying generalizability theory to the determination of the limits of generalizability in validity generalization studies. Validity generalization studies such as those conducted by Schmidt and Hunter involve the accumulation of very large distributions of validity coefficients. These validity coefficients (with the proper transformation to avoid the biased sampling distribution), can be treated as scores. The objects of measurement are not persons but studies. Facets describing the various kinds of

conditions in which validation studies are performed can be identified. An obvious example is the distinction between predictive and concurrent studies. A less obvious distinction might call for information about the subject populations and their degree of urbanization, or it might call for information about structural characteristics of the organizations. It may be that a question might arise over the generalizability of a particular validity coefficient over different industries, so that the industry of choice might be a facet. Generalizability coefficients can be computed for generalizability over various facets of interest and, by this technique, potential limits to the generalizability of a validity coefficient can be obtained. The issue of the limits of generalizability seems to be a very important one. Proper use of generalizability theory designs can lead to defined limits of the generalizability of work sample scores and of the generalizability of criterion-related validity coefficients.

REFERENCES

Brennan, R. L. Generalizability analysis: Principles and procedures. (ATC Technical Bulletin No. 26). Iowa City: American College Testing Program, 1977.

Cronbach, L. J. On the design of educational measures. In D. N. M. De Gruijter & L. J. Th. van der Kamp (Eds.) Advances in psychological and educational measurement. London: Wiley, 1976.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements. New York: Wiley, 1972.

Ghiselli, E. E. The validity of occupational aptitude tests. New York: Wiley, 1966.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. The generalizability of student ratings of instruction: Estimation of the teacher and course components. Journal of Educational Measurement, 1978, 15, 1-13.

Glaser, R., & Klaus, D. J. Proficiency measurement: Assessing human performance. In R. M. Gagne (Ed.) Psychological principles in system development. New York: Holt, Rinehart, & Winston, 1962.

Guion, R. M. Personnel testing. New York: McGraw-Hill, 1965.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement, 1976, 13, 171-183.

Lang, D. A. The work sample test: Implications for personnel research. Unpublished manuscript, 1978. (Available from Darryl Lang, Department of Psychology, Bowling Green State University, Bowling Green, Ohio, 43403.)

Lawshe, C. H. What can industrial psychology do for small business (a symposium). 2. Employee selection. Personnel Psychology, 1952, 5, 31-34.

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

Wheaton, G. R., Fingerman, P. W., & Boycan, G. G. Development of a model tank gunnery test. (Technical Report TR-78-A24). Alexandria, Va.: U.S. Army Research Institute for the Behavioral and Social Sciences, 1978.